Software Engineering Institute

Issues and Opportunities for Improving the Quality and Use of Data in the Department of Defense

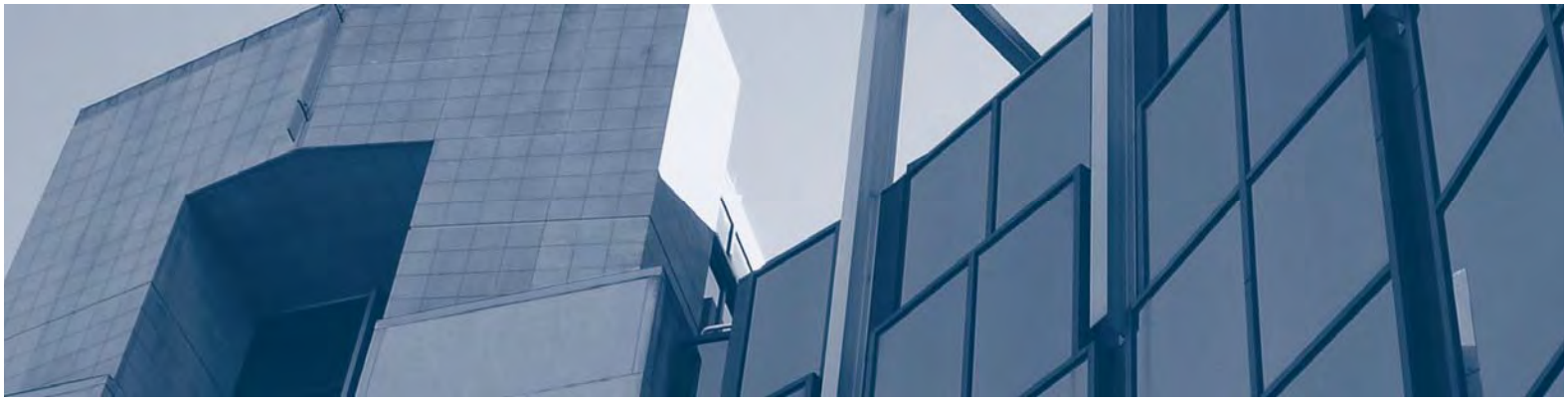
Mark Kasunic
David Zubrow
Erin Harper

March 2011

SPECIAL REPORT
CMU/SEI-2011-SR-004

Software Engineering Measurement and Analysis

http://www.sei.cmu.edu



Carnegie Mellon

This report was prepared for the

SEI Administrative Agent
ESC/XPK
5 Eglin Street
Hanscom AFB, MA 01731-2100

The ideas and findings in this report should not be construed as an official DoD position. It is published in the interest of scientific and technical information exchange.

This work is sponsored by the U.S. Department of Defense. The Software Engineering Institute is a federally funded research and development center sponsored by the U.S. Department of Defense.

This work was created in the performance of Federal Government Contract Number FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. The Government of the United States has a royalty-free government-purpose license to use, duplicate, or disclose the work, in whole or in part and in any manner, and to have or permit others to do so, for government purposes pursuant to the copyright license under the clause at 252.227-7013.

For information about SEI publications, please visit the library on the SEI website (www.sei.cmu.edu/library).

Table of Contents

Acknowledgments

The authors of this report gratefully acknowledge the participants of the workshop, whose contributions and energy made it possible. We are also grateful to Dr. Cynthia Dion-Schwarz, Director, Information Systems and Cyber Security, DDR&E, OSD-ATL, who sponsored this workshop, and Dr. Michael May, Associate Director for Software Technologies, ASD (R&E), OSD-ATL, for his review and feedback on the workshop organization and also for his participation in the workshop. We would like to thank Mr. David Jakubek, Deputy Director, Information Systems, DDR&E, OSD-ATL, for his participation in the workshop and subsequent feedback. Finally, we would like to thank Ms. Lindsay Jones, OSD-ATL for providing feedback on the workshop proposal.

Executive Summary

The Department of Defense (DoD) is becoming increasingly aware of the importance of data quality to its operations, leading to an interest in methods and techniques that can be used to determine and improve data quality. The Office of the Secretary of Defense for Acquisition, Technology, and Logistics (OSD [AT&L]), Director, Defense Research & Engineering (DDR&E) sponsored a workshop to bring together leading researchers and practitioners to identify opportunities for research that could benefit the DoD, centered around the topics of data quality, data analysis, and data use. The Software Engineering Institute (SEI) led this three-day workshop, held in Arlington, Virginia, October 26-28, 2010.

Workshop attendance was by invitation only. Seventeen papers were accepted for presentation during the first two days of the workshop. During workshop discussion, participants were asked to identify challenging areas that would address technology gaps and to discuss research ideas that would support future DoD policies and practices. On the final day of the workshop, participants worked in teams to develop proposals for further research. The SEI formed three primary recommendations for areas of further research from the five proposals produced at the workshop. These are summarized below and provided in full in Section 5 of this report.

Recommendation 1: Enable the effective integration of data from multiple and disparate data sources.

DoD information systems have been developed to respond to specific problems, with different technology solutions developed for each problem. Because there is no overriding architecture to guide the development of these systems, integrating and coordinating the data from multiple sources is difficult or impossible. Fragmentation of data across multiple systems causes unacceptable data gaps and errors due to uncoordinated data definitions, disparate quality control sets, and inconsistent subject matter vocabularies. The volume of work and the amount of time necessary to overcome these limitations is overwhelming. There is every reason to expect that developing well-integrated DoD information systems would result in important benefits.

Recommended research topics for unifying the structure of various data sources includes the development of a set of standard patterns for developing an essential architectural model, documented semantic vocabularies, and tools for using the patterns and vocabularies during system development and sustainment to ensure data quality by design. Additionally, problems integrating data could be addressed through the application of new algorithms and techniques for entity resolution.

Recommendation 2: Employ provenance analytics to ensure data quality to support mission success.

Data provenance can be used to ensure that users of data understand important background aspects, including its origin, who or what process created the data, how it was transformed, and any other conditions used to generate the data provided to users. Data provenance has the potential to be a powerful mechanism for characterizing and improving data quality. Emerging approaches to

provenance have focused mainly on metadata about the source of the data, but opportunities exist beyond this current focus.

The proposed research in this area recommends focusing on provenance analytics, including new techniques to model, record, extract, and manage provenance. In particular, techniques to reason about and use provenance should be explored, including the development of tools that use provenance information to determine data quality and how to improve it.

Recommendation 3: Develop models, methods, and tools that support data quality by design during software development.

When data quality is addressed in isolation from the development of the software used by stakeholders to collect, process, analyze, and report it, the value and benefits of the data and the enterprise's resultant information products can suffer. Missed opportunities, bad analysis, and incorrect decisions can all result from a failure to integrate data quality with software requirements. Software developers may assume that data quality considerations have been addressed by other stakeholders such as business analysts or data architects, or they may make wrong assumptions regarding data quality and associated requirements.

The research proposed is an exploration of ways to formally address data quality as part of the software development life cycle. Specifically noted was the need for a data quality calculus to provide an analytical method for characterizing the degree of uncertainty in information products based on the quality of data used to produce them. The data quality is presumed to be a function of the extent to which it was explicitly addressed as part of system development.

Abstract

The Department of Defense (DoD) is becoming increasingly aware of the importance of data quality to its operations, leading to an interest in methods and techniques that can be used to determine and improve the quality of its data. The Office of the Secretary of Defense for Acquisition, Technology, and Logistics (OSD [AT&L]), Director, Defense Research & Engineering (DDR&E) sponsored a workshop to bring together leading researchers and practitioners to identify opportunities for research focused on data quality, data analysis, and data use. Seventeen papers were accepted for presentation during the workshop. During workshop discussion participants were asked to identify challenging areas that would address technology gaps and to discuss research ideas that would support future DoD policies and practices. The Software Engineering Institute formed three primary recommendations for areas of further research from the information produced at the workshop. These areas were integrating data from disparate sources, employing provenance analytics, and developing models, methods, and tools that support data quality by design.

1 Introduction

1.1 Background

The Department of Defense (DoD) is becoming increasingly aware of the importance of data quality to its operations. The negative impacts of poor quality data have been acknowledged in several recent studies and government reports. Poor data costs the DoD and federal government billions of dollars per year: an estimated $13 billion for the DoD and $700 billion for the federal government [English 2009].

Government Accountability Office (GAO) reports indicate the DoD also suffers from the use of poor quality data in its measurement and analysis infrastructure [GAO 2009]. The consequences of this can be significant:

- Data and information reported to program offices is often flawed or incomplete. This data is further reported to oversight organizations, stored in repositories, and used as input for future evaluations and decisions.
- Flawed data negatively influences the development of acquisition policy and processes.

This growing recognition of the need to improve data quality is fueling more efforts to understand how data should be collected, analyzed, stored, and shared.

1.2 Workshop Overview

A three-day workshop was held at the Software Engineering Institute in Arlington, Virginia, October 26-28, 2010. The workshop was sponsored by the Office of the Secretary of Defense for Acquisition, Technology, and Logistics (OSD [AT&L]), Director, Defense Research & Engineering (DDR&E).

The goals of the workshop were to

- bring together leading researchers and practitioners to identify rich opportunities for research focused on data quality, data analysis, and data use
- identify the *challenging* areas that will address technology gaps, both short term and long term
- identify research ideas and technologies to support future policies and practices focused on improving the quality and value of data within the DoD
- promote further research on the above topics

Workshop attendance was by invitation only. Seventeen papers were accepted for presentation during the first two days of the workshop. On the final day of the workshop, participants worked in teams to develop several proposals for further research, using Heilmeier's Catechism to structure their efforts [Heilmeier 1991].

1.3 Workshop Participants

Presenters

Dr. Nabil R. Adam	U.S. Department of Homeland Security
Dr. Peter Aiken	Data Blueprint, Inc.
Dr. Diana I. Angelis	Naval Postgraduate School
Dr. Mary Maureen Brown	University of North Carolina at Charlotte
Dr. Susan B. Davidson	University of Pennsylvania
Mr. Brett Dorr	DataFlux
Mr. Robert Flowe	OUSD (AT&L)
Mr. David C. Hay	Essential Strategies, Inc.
Mr. John Horst	National Institute of Standards and Technology (NIST)
Dr. Alan Karr	National Institute of Statistical Sciences (NISS)
Mr. David Loshin	Knowledge Integrity, Inc.
Dr. Douglas J. MacKinnon	Naval Postgraduate School
Dr. Tim Menzies	West Virginia University
Dr. Sudha Ram	University of Arizona
Dr. Thomas C. Redman	Navesink Consulting Group, LLC
Mr. David Schlesinger	Metadata Security, LLC
Dr. John R. Talburt	University of Arkansas at Little Rock

Attendees

Dr. Jon Agre	Institute for Defense Analyses
Ms. Anita Carleton	Software Engineering Institute
Ms. Erin Harper	Software Engineering Institute
Mr. David Jakubek	OUSD (AT&L), DDR&E
Mr. Mark Kasunic	Software Engineering Institute
Ms. Patricia Lothrop	OUSD (AT&L)
Dr. Michael May	OUSD (AT&L), DDR&E
Mr. Ryan Mckenzie	Amentra, Contract Support to OUSD (AT&L)
Dr. Richard Wang	Headquarters, Dept. of the Army
Dr. David Zubrow	Software Engineering Institute

1.4 About This Report

Section 2 of this report summarizes the papers presented at the workshop. Section 3 describes the results of two brainstorming activities that were conducted to set the context for the discussion of data quality problems and to identify possible innovations. Section 4 lists the research proposals

that were developed by groups of workshop participants. Section 5 lists three recommendations that the SEI identified for further research based on the proposals in Section 4. The appendix contains a copy of the agenda used for the workshop.

2 Workshop Presentations

This section contains summaries of the papers presented at the workshop. The papers and the paper summaries reflect the views of the authors, not those of the SEI or the workshop sponsors. Most of the papers are available in full on the SEI website at http://www.sei.cmu.edu/measurement/research/dataworkshop.cfm. For papers not available through the site or for questions about the papers' contents, please contact the authors.

Presentations were grouped into the following six topics:

1. Data quality issues in DoD acquisition
2. Managing data quality
3. Understanding the quality of real-time data
4. Data quality monitoring and impact analysis
5. The role of data provenance
6. Modeling data quality

2.1 Data Quality Issues in DoD Acquisition

OUSD AT&L Perspectives and Initiatives: Acquisition Visibility and Data Quality

Robert Flowe

Robert Flowe from the Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics (OUSD AT&L) gave an introductory presentation about data quality needs in the DoD. The DoD uses acquisition data for management and oversight. The Acquisition Visibility (AV) initiative was established to improve the use and availability of the data, specifically to (1) make acquisition data more visible and useful, and (2) make it easier on the people required to report data.

The AV initiative has three themes:

- Data governance – Data governance pertains to the laws, regulations, and policies that govern data generation. There are no widely-employed data standards in use. With many local policies in place, opportunities need to be indentified to streamline and de-conflict these policies to reduce redundancy and improve data quality.
- Data transparency – The reliability of data needs to be enhanced through monitoring and validation. The ability to link to authoritative data from a responsible source helps answer questions about reliability, trust, and provenance. A goal is to have a "one-stop shopping" place with a single entry point where DoD senior leadership can get the best data.
- Value-added data services and tools – Functionally-tailored tools that use authoritative data are needed to support DoD acquisition management and oversight. These tools could help to address problems with business rules, exception handling, validation, and estimation. Further

research is needed to determine where automation or augmentation is appropriate. These tools could provide the ability to check data quality on the way in and on the way out and determine if algorithms are being applied correctly.

Flowe outlined some challenges to getting good data, including capturing business rules and other metadata, validating correctness, and handling exceptions. Better data quality tools and methods could help meet these challenges. Specifically, methods are needed for assessing the maturity of a data source, handling the subtle business rules that provide necessary context to the data, and assessing data quality at the "wholesale" versus "retail" level. Flowe concluded by proposing a new framework for a data quality model.

Resource Management Decision (RMD) 700 and Programmatic Interdependencies

Mary Maureen Brown

In her paper, Mary Maureen Brown provided a brief overview of ongoing research that seeks to isolate the programmatic success factors of joint capability initiatives. A data shortfall recognized during this research and also observed by AT&L's Acquisition Visibility Service-Oriented Architecture (AV SOA) initiative led to the issuance of Resource Management Decision directive 700 (RMD 700), which seeks to improve major defense acquisition program (MDAP) resource data transparency.

Brown pointed out that although program interdependencies are growing rapidly, little attention has been devoted to understanding the nature of the interdependencies and their potential consequences. The research she described examines DoD acquisition "from the context of a network of interrelated programs that exchange and share resources for the purpose of establishing joint capabilities." The goal is to determine which resources are being exchanged as part of the interdependencies.

When attempting to examine budget exhibits in the context of interdependencies, Brown said, it became evident that the cost data among various reports were incomplete, inconsistent, and largely inaccurate. As a consequence, it was impossible to isolate the true cost of a given MDAP initiative.

This contributed to a consensus that there was a need for a greater focus on data standards, data transparency, and data visibility. Spearheaded by the AV SOA initiative, in partnership with the Comptroller and Cost Assessment and Program Evaluation (CAPE), RMD 700 seeks to address the issue of cost data quality.

While discussing future research opportunities, Brown pointed out that 95% of the MDAPs have problems in their cost, procurement, and budget data. To move into the new world of joint capabilities, Brown concluded, we need to ask the research questions that will tell us how to mitigate these problems.

Further areas of research identified by Brown include the following:

- Behavioral changes
- Policy changes (i.e., Resource Management Decision Directive 700)
- Technical needs, such as

- interoperability in bringing various data sets together and maintaining them while they are in constant change
- multidimensional scaling, or the ability to look at problems "on the fly" in multiple dimensions
- agent-based systems
- pattern recognition
- visualization, or the ability to show data in a way that helps people understand it and make decisions

Measuring Transaction Costs in DoD Acquisition Programs

Diana I. Angelis, John Dillard, Raymond Franck, and Francois Melese

Conventional cost estimation techniques used to support the acquisition of major weapon systems rely on production costs, and as a result most of the data in DoD cost databases also relate to production costs. Diana Angelis stated that transaction cost economics "suggest that another set of costs—transaction costs (the costs of coordination and motivation) should also be considered in developing a cost estimate." She noted that the first step in doing so is to identify and measure transaction costs in DoD acquisitions.

Angelis went on to describe research efforts undertaken to measure transaction costs in DoD acquisition programs. Two research projects, sponsored by the Acquisition Research Program at the Naval Postgraduate School, were conducted by Diana Angelis, John Dillard, Chip Franck, and Francois Melese. This research "suggests how transaction costs for DoD acquisitions might be measured and discusses difficulties encountered in using data in DoD cost databases to support the measurement of transaction costs for major weapon systems."

Angelis stated that after examining the data, the need for standardization in the information collected for cost databases became readily apparent. She suggested that "for a given category, the data collected should be consistent over time and across categories. If there are necessary changes in the data categories, there should be a mapping from the old categories to the new categories. Differences between databases should be reconciled to provide better quality information to decision makers and cost analysts."

Angelis stated that measuring transaction costs in DoD acquisition may improve cost estimating for major weapon systems, but current practices do not support measuring these costs, directly or indirectly. In conclusion, she said that "measuring and reporting transaction costs may lead to more complete cost estimates and provide the DoD with important information about the effectiveness and efficiency of the acquisition process."

2.2 Managing Data Quality

Practical Considerations for Rapidly Improving Quality in Large Data Collections

Peter Aiken

In his paper, Peter Aiken discussed the fundamental issues in existing approaches to improving DoD data quality. A major problem noted by Aiken is that most organizations are still approaching data quality problems from stove-piped perspectives. He also pointed out that because many data quality challenges are unique or context specific, the resolution of those challenges cannot

follow programmatic solution development practices. Therefore, the development of "data quality engineering specialists" in organizations is needed. Also contributing to the problem, according to Aiken, is that educational institutions are not addressing the issues and vendors "are incented to not address the challenges proactively."

Aiken also described possible solutions to many of these problems. First he pointed out that a body of knowledge (BoK) has already been developed: the Data Management BoK published by DAMA International. The BoK describes primary data management functions focused on delivering data to the organization, structured around several environmental elements.

Improvements to available tool sets could also contribute to solutions, and a multitude of products are now available to assist with various analyses and tasks. Aiken proposed that "the most common problem now facing the DoD is the widespread perception that tools alone will accomplish data quality improvements and that the purchase of a software package can solve data quality problems." Aiken further stated that "the best data quality engineering solutions will continue to be a combination of selected tools and specific analysis tasks, with the primary challenge in determining the proper mix of human and automated solutions."Aiken concluded his paper with the idea that data quality must be approached as a specialized discipline.

Issues and Opportunities for Improving the Quality and Use of Data Within the DoD

David Hay

In his paper, David Hay discussed how the DoD can ensure that the right people get the right data when they need it. He focused on the complexity of data structures and proposed that "the true, underlying structure of the concepts described by the DoD's data is considerably simpler than most of their embodiments in real databases." Hay explained that understanding the simpler underlying structure is essential in making sense of the overwhelming complexity of real databases and systems.

Hay went on to say that data architecture and data presentation require different orientations, and that data are most effectively managed so that as business situations change, these changes can be addressed with changes to configuration data values—without having to change the underlying structure of a model or database. Data presentation, in contrast, must be in terms the customer understands. Customers have particular "views" of the data, which tend to be in the concrete terms of their daily experience.

Hay concluded by outlining three kinds of improvements that he considers necessary for better data quality:

1. Improved processes – Data capture can be made intuitive, and required data checks can be built in to make them impossible to violate.
2. Improved data models – If the data are badly organized, with duplicates and poorly related tables, then errors are easier to make. Also, if customers do not understand the underlying structure of the data, they might formulate queries that return something very different from what they expected or completely wrong. Some problems occur because the data model did not correctly reflect the underlying structure of the world; others because the customer did not understand the implications of the model.

3. Improved training – Training must be an ongoing process. Customers need to be taught the implications of the data structure in their data model, and the people responsible for capturing data need to be taught the proper way to do so.

Taxonomy to Drive Data Quality, Security, and Regulatory Compliance in a Global Enterprise

David Schlesinger

Metadata-driven data management uses multi-dimensional metadata to manage data security and define the data flow through the enterprise. In his paper, David Schlesinger discussed the use of metadata-driven management in the context of taxonomy for metadata that can drive data quality and protection.

Schlesinger suggested that the issues of data quality and data security "will not be managed maturely until we accomplish three interim goals:

1. We must abandon our traditional ideas of what constitutes information definition and expand it to include all attributes important to the organization.
2. Traditional security classification structure must evolve to allow a richer data description that includes metadata characteristics for quality capability, regulatory compliance, and data protection.
3. There must be a centrally accessible and authoritative data dictionary (i.e., metadata repository) that holds the security, regulatory, and quality definitions for every type of information used within the enterprise."

Schlesinger concludes that "regulatory compliance, data security, and accurate entitlement decisions require access to rich metadata resources, crisp definitions, and documented procedures, and must extend across the organization uniformly to provide this governance knowledge as a service to all business and operational units."

Security and Confidentiality in Outsourced Databases

Nabil Adam, Bijit Hore, and Sharad Mehrotra

In this paper, Nabil Adam discusses the emerging paradigm of outsourcing databases. He related that outsourcing provides significant advantages to users, including reliability, continuous availability, accessibility from many locations, and performance at a fraction of the cost and effort compared to creating a similar service in-house.

He also said that there are significant risks to outsourcing. The data resides outside the control of the user, which can pose significant security and privacy risks. Adam points out that recent research has addressed many of the security challenges that arise in the outsourcing model, but a complete solution to outsourcing that offers efficient storage, retrieval, sharing, and mobile access but also ensures data privacy and confidentiality remains a challenge.

He states that it is unlikely that all security vulnerabilities in the outsourcing model can be completely eliminated using specialized data hiding or encryption-based strategies. Instead of trying to completely eliminate disclosure risks, Adam proposes a new direction of research that explores strategies to contain and manage risks in the outsourcing solutions. This approach would identify the inference channels that might compromise data confidentiality, characterize their interaction

and interdependence, and then reduce the overall risk of exposure instead of completely eliminating it.

Another fundamental challenge described by Adam was support for secure information sharing in the outsourced model. While the ability to seamlessly share data with others has helped popularize existing outsourcing solutions, little research has been done to explore the challenge of supporting secure data sharing. Adam concluded with a discussion of an approach to privacy and data confidentiality using a privacy middleware that sits between the client and server and implements the client's privacy policies on the outgoing data.

Issues and Opportunities for Improving the Quality and Use of Data within the DoD

Douglas MacKinnon, Ying Zhao, and Shelley Gallup

Douglas MacKinnon described several challenges and motivations related to the DoD's efforts to improve data quality and use. The first challenge he identified relates to the source of the data: DoD applications often include data from disparate real-time sensor and archival sources with multiple dimensions. The high rates of collection for these kinds of data mean that large volumes accumulate quickly. While some of the data collected is structured data in traditional forms, a large percentage of the data is unstructured (e.g., in free text, Microsoft Word documents, PDF documents, PowerPoint documents, and emails). He pointed out that providing for the storage, analysis, search, and retrieval of the data in each source while at the same time providing a cross-examination of all the data to create a full picture—a form of "situation awareness"—is a daunting task. Analysts need automated analysis tools to create and sustain situational awareness in real time.

Another problem is that intelligence data "arrives from myriad sources that overwhelm analysts' abilities to perform the necessary deeper intelligence analyses that result in timely, situational awareness." Currently, analysts "manually comb through immense volumes of multi-source, multi-classification level intelligence data to find previously unknown, undiscovered patterns, associations, relationships, trends, and anomalies."

MacKinnon described current research that applies a data-driven automation system called lexical link analysis (LLA) to help DoD researchers and decision makers "recognize important connections (concepts) that are patterns derived from dynamic, ongoing large volumes of DoD data collection." Lexical analysis is a form of text mining in which word meanings represented in lexical terms are developed from their context, while link analysis is a subset of network analysis that explores associations between objects and reveals crucial and sometimes unexpected relationships. LLA is an extension of lexical analysis combined with link analysis.

MacKinnon asserted that if this research is successful, it could facilitate real-time awareness and reduce the workload of decision makers. He listed the following areas he believes will be impacted by this research:

- understanding and managing massive, dynamic data
- analysis of social, cultural, and linguistic data
- multidisciplinary approaches to assessing linguistic data sets
- extraction and representation of the information in non-technical, unstructured documents

- effective analysis of massive, unreliable, and diverse data
- analysis of significant societal events

2.3 Understanding the Quality of Real-Time Data

Achieving Information Quality

John Horst

In his paper, John Horst said that while the storage of raw data still happens, it is increasingly being replaced by data stored in an organized and meaningful structure. This "data plus meaning" is what Horst describes as information quality, and he asserted that "achieving information quality is even more challenging than achieving data quality."

One problem related to data quality, Horst said, is that data are often presented to the consumer without associated descriptive semantics. These semantics and summary statistics, however, "are essential to the efficient perception and application of data content to process improvement." He suggested that this problem could be mitigated by refining guidelines for defining and associating semantics and summary statistics to raw data. He also advocated working backward and looking at common processes and determining how they dictate data storage and what additional semantic information is required. With software versions and data formats changing frequently, it is also essential to define standard procedures for continuous data validation and verification.

Horst stated that communication is essential when going beyond data quality improvements to improving information quality, and information quality can be degraded by a failure to communicate. The failure commonly occurs because there is a proliferation of languages for the same information. Translation is required, leading to information quality losses and other significant costs. To mitigate this problem and maximize information (and data) quality, he recommended the use of well-defined languages and uniform implementations of these languages by information generators and information consumers.

Information Quality Education and Research at UALR

John Talburt

John Talburt, advisor for the Master of Science in Information Quality (MSIQ) program at the University of Arkansas at Little Rock (UALR), provided an introduction to the education and research initiatives at the university and their potential for improving the quality and use of DoD data. His paper also described the three tracks of current DoD-sponsored research at the university:

1. Identify information quality requirements for the layered sensor domain. This project focuses on the identification of information quality metrics related to the analysis of video data streams.
2. Prototype interactive 3-D information visualization in the layered sensor domain. This project focuses on using interactive 3-D visualization to improve the quality of information available in the layered sensor domain.
3. Create visual rendering and display of text and information quality metrics. This project proposes that integrated processing of data from multiple types of sensors can benefit a variety of decision making processes. An information processing scheme is being prototyped that of-

fers data fusion for multiple sensors. For example, data from temperature sensors or motion detectors could be fused with data from visual sensors, such as security cameras.

Talburt pointed out that, in addition to research, the UALR graduate program also provide education and training opportunities that could benefit the DoD. One of the goals of the UALR Information Quality graduate program is to help establish IQ as a recognized academic discipline, and the school is the only university in the U.S. to offer graduate-level degrees in information quality.

Talburt concluded with a description of other related research topics being explored at UALR:

1. The OYSTER project - OYSTER (Open sYSTem Entity Resolution) is an open-source entity resolution system based on the identity capture architecture that supports probabilistic direct matching, transitive linking, and asserted linking. It also supports persistent identity identifiers.
2. Information hubs – entity resolution, particularly identity resolution, is proving to be a critical process in the movement to create systems that facilitate information sharing across entity-centric, inter-organizational data stores. An emerging solution to this problem is to build a single system that connects to each of the participating organizations.

2.4 Data Quality Monitoring and Impact Analysis

Evaluating the Business Impacts of Poor Data Quality

David Loshin

In his paper, David Loshin described types of risks attributable to poor data quality and presented an approach to correlating business impacts to data flaws. He began by examining ways to determine the value of information and suggested that understanding the utility—or expected value to be derived from the information—is the best approach to determining value. Analyzing how data is being used to achieve business objectives and how flawed data can impede those goals is important because data quality is subjective. To determine the value added when the quality of the data is improved, the data's role in meeting business expectations must be understood. This involves identifying business impacts, related data issues, and their root causes; then quantifying the costs to eliminate the root causes.

Loshin said that to analyze the degree to which poor data quality impedes business success, a classification scheme can be used to categorize business impacts associated with data errors. The classification scheme he identified uses a simple taxonomy with four primary categories that relate to the negative impacts of data errors or the potential business improvements that could result from improved data quality. The four categories he mentioned include

- financial impacts
- confidence- and satisfaction-based impacts
- productivity impacts
- risk and compliance impacts

Loshin emphasized that the four high-level categories he identifies are not necessarily exhaustive. Since every industry, business, and organization is different, identifying impact requires an examination of how data is being used to achieve business objectives. He noted that "classifying the

business impacts helps to identify discrete issues and relate business value to high quality data, so it is valuable to examine these impact categories closely. Impacts can be assessed within enumerated subcategories, which help refine a means for quantification."

According to Loshin, quantifying and establishing a hierarchy for data quality issues provides two primary benefits. First, classifying impacts in small increments makes the determination of the impacts of poor data quality more manageable. Second, the categorical hierarchy of impact areas maps to a performance reporting structure for gauging improvement. Identifying where poor data quality impacts the business and where data quality improvement will help the business provides a solid framework for quantifying measurable performance metrics that can be used to develop key data quality performance indicators.

Data Quality, Logistics, and USTRANSCOM

Brett Dorr

In his paper, Brett Dorr of DataFlux Solutions discussed the role of data in the logistics process. DataFlux and other organizations have provided assistance to the United Transportation Command (USTRANSCOM)—the group that provides air, land, and sea transportation for the Department of Defense—in their efforts to create a more effective data quality and data monitoring program.

Dorr first identified two primary data elements that can cause problems in logistics programs:

1. Improper or inadequate location information. Destination data that is missing, inaccurate, or unreliable can cause shipments to be delayed significantly or to arrive at the wrong destination.
2. Inability to track and manage complex shipments.

According to Dorr, many problems related to the first element have been prevented or mitigated through the implementation of data quality and validation efforts over the last 20 years. The second element – which requires understanding and managing data on materials or products – is inherently more problematic. Standards for product data are not as codified as those for customer data. Product data has conventions, but they are not as universally defined and understood as those for customers.

Applying logic to data can help with problems related to this element. Dorr noted that one approach is to create a way for information technology systems to replicate the same logic (i.e., business rules) that are already part of the intellectual fabric of the organization. To do this, the rules cannot be linked to any specific data source or database table, but instead mapped to any data source, text file, or real-time web service. This allows an enterprise to administer a single set of rules across all database systems and application transactions in the organization, decreasing the cost of implementing the quality rules.

After the business rules are established, the next step is to enable continuous monitoring of the data. Dorr related that this is "best accomplished by setting up a data monitoring repository, which stores information about events or 'triggers' that require attention within the system." After monitoring business rules for a period of time, organizations can use the information they gather as the foundation for reports on data trends.

This method was applied to a problem faced by USTRANSCOM: the need to check the Transportation Control Numbers (TCNs) assigned to items or pallets. USTRANSCOM used DataFlux technology to monitor these codes as they passed through the system. When the system detected an invalid TCN, the system alerted the appropriate employee. A new level of intelligence was also provided that enabled the ability to "teach" the DataFlux software how to check for a variety of values related to the TCN.

This project demonstrates a successful example of using data monitoring technology to enforce business rules that were already established—in this case, for the logistics process. Dorr suggested that other organizations could use a similar approach to better integrate existing logic into information technology systems by creating and managing rules in an automated way.

2.5 The Role of Data Provenance

Data Provenance: The Foundation of Data Quality
Susan Davidson and Peter Buneman

Provenance is fundamental to understanding data quality. In her paper, Susan Davidson explored provenance issues related to the creation, transformation, and copying of data. She stated that models of provenance need to be developed that take these issues into account and described two general models of provenance that have been developed: workflow provenance and data provenance.

Davidson asserted that scientific data sets are seldom "raw." Instead, they are the result of a sophisticated workflow that takes raw observations and processes them using a complex set of data transformations and possibly input from a scientist. Workflow provenance represents a record of what was done to transform these observations into scientific datasets. Data provenance is more fine-grained and is concerned with relatively small pieces of data instead of an entire system. Davidson said that a convergence of the models for workflow provenance and data provenance is needed.

However, ensuring provenance is captured in the first place is required before these models can be used. The current state of the practice for capturing provenance is manual—that is, humans enter the information. Davidson described three advances being made in the area of provenance capture:

1. Schema development - Several groups are reaching agreement on what provenance information should be recorded, whether through manual or automatic capture. Various government organizations are defining the types of metadata that should be captured.
2. Automated capture - Some scientific workflow systems automatically capture processing steps and their input and output data. This "log data" is typically stored as an XML file or put in a relational database to enable users to view provenance information. Other projects are focusing on the capture of provenance data at the operating system level.
3. Interception - Provenance information can be "intercepted" at observation points in a system. For example, copy/paste operations could be intercepted to record the location of a piece of data being copied. Another strategy is to capture calls at multi-system "coordination points."

Davidson related that privacy also presents challenges in capturing provenance: making complete provenance information available to all users can raise privacy concerns. For example, intermediate data in a workflow execution may contain sensitive information, or a module itself may be proprietary, meaning that users should not be able to infer its behavior. While some systems attempt to hide the meaning or behavior of a module by hiding its name or source code, this does not work when provenance is revealed. These kinds of issues point to a tradeoff between the amount of provenance information that can be revealed and the privacy guarantees of the components involved.

Archiving evolving data sets is also a challenge, according to Davidson. Although tools have been developed for space-efficient archiving, few publishers of online data do a good job of keeping archives. This means that the provenance trail can be lost even if people keep provenance information. A related issue is the need for keeping data citations.

Finally, Davidson listed a number of problems that will require further understanding, including

- operating across heterogeneous models of provenance
- capturing provenance
- compressing provenance
- securing and verifying provenance
- efficiently searching and querying provenance
- reducing provenance overload
- respecting privacy while revealing provenance
- using provenance for evolving data sets

She concluded that while more research is needed in this area, the major challenges to capturing provenance "are in engineering the next generation of programming environments and user interfaces and in changing the mind-set of the publishers of data to recognize the importance of provenance."

Determining Data Quality Based on Provenance

Sudha Ram and Jun Liu

In her paper, Sudha Ram related that "one of the most important uses of data provenance is to estimate data quality and reliability. However, little research has been done on how data provenance can be used to assess the quality of the data." Ram then described two studies conducted to assess the quality of data based on its provenance.

The first study used the provenance of Wikipedia articles to classify contributors based on their roles in editing articles and to identify various patterns of collaboration. The intent of the study was to help identify collaboration patterns that are preferable or detrimental for data quality. The second study, being done in collaboration with the defense industry, involves the development of a data quality framework that uses data provenance to automatically assign a quality grade to material data.

The Wikipedia study began with the development of a domain ontology based on the W7 model, which is an ontology that conceptualizes data provenance as consisting of seven interconnected

elements: what, when, where, who, how, which, and why (7 W's). Next, the provenance was extracted and the various roles played by contributors for a given article were identified. Several collaboration patterns were uncovered, each characterizing a distinctive way in which a group of contributors collaborated. Finally, the quality of the articles was examined and statistical methods were used to determine the impact of the collaboration patterns on the quality of the Wikipedia articles.

The second study was motivated by the DoD's need to understand the quality of data about materials. In DoD projects, the use of low quality data may result in poor or improper material selection, with serious consequences. However, material data often look alike. Ram related that the provenance of material data, such as who provided the data and how the data was generated, is critical knowledge that helps material engineers discern low quality data from high quality data. A domain adaption of the W7 model is also being used for this study, and a subset of the model (including what, how, who, and when) is being applied to track the provenance of material data.

Both of these studies provide insight into the systematic use of provenance in assessing data quality. Ram noted that future plans include (1) developing innovative approaches to tracking provenance, and (2) automatically assessing the quality of data available in the DoD product development industry.

2.6 Modeling Data Quality

Selecting Quality Data

Tim Menzies

In his paper, Tim Menzies presented the results of two studies that demonstrate the use of key selection operators to remove unnecessary and noisy data, exposing significant information. Because collecting data is expensive and time-consuming, he said, it may seem strange to suggest that some data be removed and discarded. However, the studies in his paper show that throwing away parts of the data can led to significant improvements.

Menzies lists several reasons that data should be discarded:

- Noise – If data collection is flawed, one or more features of the data may be noisy (i.e., contain spurious signals not associated with variations to projects).
- Correlated features – If multiple features are tightly correlated, then using all of them will diminish the likelihood that either feature attains significance. A repeated result in data mining is that removing some of the correlated features increases the effectiveness of the learned model.
- Under-sampling – The number of possible influences on a project is large, and historical data sets on projects for a particular company are usually small. This motivates the removal of features, even if those features are theoretically useful.

In data mining, simpler models with equivalent or higher performance can be built using feature subset selection algorithms that intelligently remove useless features.

In the first case study, the WRAPPER algorithm developed by Ron Kohavi and George John was applied to COCOMO effort estimation data. Using WRAPPER, over 65% of the columns in the

data set were pruned. This pruning was shown to improve estimation effectiveness, sometimes dramatically. For data sets with less than 20 examples, the "after" results were better by a factor of 2 to 4. With one exception, the general trend indicates that as data set size shrinks, the improvement increases. That is, pruning is most important when dealing with small data sets.

In the second case study, a feature selector for significance was applied to Bayes networks that represented expert intuitions about the connections of factors that degrade software quality. This was done to determine which nodes in these networks were most influential in creating the worst defects. Menzies listed two important conclusions:

1. When selecting for the worst defects, only a few ranges are important.
2. Most features have nearly zero significance (i.e., most features do not matter).

Menzies said that for anyone writing a software data quality plan, the results of these studies have major implications: it may be "a waste to expend great effort to precisely define all possible data points, and then spend time collecting data according to those definitions."

He concluded that while definition and collection plans for data quality are important, "the right level of data quality may be lower than currently suspected. Selection operators can repair low-quality data, at least to the point where significant business decisions can be made about the project. Hence, when writing quality plans for data definition and collection, auditing those plans with an exploratory data significance study is recommended. Elaborations on the data definitions and data collection become superfluous when those elaborations no longer improve the recommendations coming from the significance studies."

Data Quality Research That Builds on Data Confidentiality

Alan Karr

In his paper, Alan Karr discussed the competing demands of providing high quality, high fidelity data for making decisions and conducting research with the necessity of protecting the confidentiality and privacy of those providing the data. His paper presented a theoretical model of the trade-off between data quality and its impact on decision making. Currently, agencies collecting statistical data alter it before releasing it for analysis, ideally in ways that substantially decrease risk with only a minor decrease in utility.

Karr lists several techniques used to mask the data, including

- dropping explicit identifiers
- suppressing cells in tables
- coarsening values
- microaggregation
- adding noise
- swapping data
- using synthetic data

He concluded by proposing a cost model for quantifying and understanding the impact of modifications to data quality on the decisions that the data inform.

3 Workshop Activity Results

On the final day of the workshop, two brainstorming activities were conducted. The goal of the first activity was to set the context for data quality research in today's environment. The goal of the second activity was to identify innovations impacting data quality, data analysis, and data use. The results of these activities were used to generate workshop findings, as discussed in Section 4.

3.1 Activity 1: Context for Data Quality Research in Today's Environment

This activity was conducted to facilitate discussion about the current landscape of data quality research. The dialogue that occurred during this activity provided a unique opportunity to examine assumptions and appreciate the wealth of information that the group had to offer. The activity allowed the group to

- develop a shared big-picture view of the environment
- lay the groundwork for breakthrough thinking
- increase understanding of some of the complexity that impacts progress in data quality
- examine assumptions and multiple perspectives
- identify factors and trends that require tracking
- establish a common backdrop for the detailed work of proposing research ideas for data quality, data analysis, and data use

The results of this exercise are summarized below, with some context added for terms that were discussed but not fully elaborated.

Department of Defense (DoD) Trends

Several trends affecting the Department of Defense were identified and their relationship to data quality discussed. The major themes are described below.

- **Increase in threats from cyber sources**

  Recent reports state that an average of 1.8 billion cyber attacks occur per month. NATO has increased its cyber security and defense efforts and has suggested that online warfare presents as serious a threat as missile attacks.[1] Detecting cyber attacks in networked information systems means analyzing very large amounts of data generated from network traffic monitoring. High performance data mining algorithms and tools are needed to support the analysis of these massive data sets.

- **More centralization of services, affecting how data is created and stored**

  Centralizing services can be mean integrating data stored across multiple databases and platforms. Moving from silos of information to a single, shared resource creates a multitude of integration problems.

[1] http://www.defensesystems.com/Articles/2010/04/26/Digital-Conflict-Cyber-Defense.aspx

- **Greater demands on resources and pressure to do more with less**

  Recent DoD efforts aim to significantly speed the procurement of information systems and also cut large amounts from the defense budget. With less money to use fixing problems, good quality data are more important than ever to the decision makers charged with implementing these efforts.

- **An increasing need for real-time data**

  Real-time data refers to data delivered immediately after collection with no delay in the timeliness of the information provided. The use of real-time data is important in sensor-based applications and in cybersecurity efforts related to monitoring network activity. However, real-time data presents unique challenges for data quality. As described in *Data Management:* "In the middle of a process [in batch-oriented data integration], you've got a chance to actually analyze and cleanse that data. In the world of real-time data integration, there's less opportunity to apply very sophisticated files for analyzing the quality and cleansing the data. There is a higher risk, then, that data integrated in real time will be of poorer quality, incorrect, or misleading."[2]

Technology Trends

Emerging technology trends and their impact on data quality were discussed by workshop participants. Several trends that were identified are described below.

- **Cloud computing**

  Cloud computing involves sharing resources, software, and information through the internet, on demand, to computers or other devices. In this shared infrastructure environment, data governance and data quality are topics of great concern.

- **High-performance computing**

  High-performance computing refers to the use of supercomputers or computer clusters to solve advanced computation problems. High-performance computing is especially important for large-scale data mining applications because of the huge data sets and the amount of computation involved.

- **Web 2.0/Semantic web technologies**

  The article “Taxonomies and the Semantic Web in Product Master Data” describes the semantic web as “an evolving development of the World Wide Web in which the meaning (semantics) of information and services on the web is defined, making it possible for the web to ‘understand’ and satisfy the requests of people and machines to use the web content” [Power 2010]. The semantic web standardizes formats for data, making them discoverable or learnable by tools. The interchange, distribution, and reuse of data can also be greatly facilitated by the infrastructures of the semantic web.

[2] http://searchdatamanagement.techtarget.com/news/1297863/Companies-choosing-real-time-data-integration-over-batch-oriented-techniques

- **Service-oriented architecture (SOA)**

  A SOA provides a set of interoperable services that can be used with multiple, separate systems. The underlying data in a SOA needs to be of high quality and organized in a common way to make it consistent across services. As reported in *eBiz*, "If data is outdated, unqualified, or duplicated, SOA-enabled services may be delivering dirty data even faster and to more users than before."[3]

Data from New or Expanded Sources

Data growth has been a major trend of the last decade, with huge collections of data being created daily. The current amount of data available is projected to grow 44 times in the next eight years.[4] All this new data brings not only new opportunities, but also challenges related to storage, accessibility, and control.

- **Social media and subject unaware data collection**

  Social media sites like Facebook and Twitter are large contributors to the explosive growth of data. Research projects are underway to help people understand how to use this unstructured, real-time data in predictive analytics, and the tools needed to do so efficiently and effectively are also being developed.

- **Proliferation of unstructured data**

  Unstructured data refers to information that is not based on a data model or structured in a way that can be easily understood by a computer program. Merrill Lynch estimates that more than 85 percent of all business information is unstructured data, including information in formats such as e-mails, memos, user groups, chats, reports, letters, surveys, white papers, marketing material, research, presentations, and web pages.[5]

  The tools and techniques used to transform structured data into useable information do not work on unstructured data. Data mining, text analytics, and noisy text analytics techniques are some methods being investigated to find patterns in or interpret this data.

Need for experts/leadership in this area

As people become more aware of the importance of data and information quality, there is an increasing demand for experts to advise government, industry, and non-profit organizations. University programs are being created to meet this need. Increased interest has also led to a proliferation of unsubstantiated analysis methods and technologies. Experts are crucial in determining which are valid and which are not.

---

[3] http://www.informatica.com/solutions/soa/Pages/index.aspx

[4] http://clarisnetworks.com/Blog/May/The-Data-Crisis--Keeping-Your-Data-Volumes-Managea

[5] http://www.information-management.com/issues/20030201/6287-1.html

3.2 Activity 2: Waves of Innovation Impacting Data Quality, Data Analysis, and Data Use

This activity was a forward-looking environmental scan of factors that may impact data quality, data analysis, and data use. The discussion provided a way to gather the combined knowledge and perspectives of the experts about areas that require further research and to identify knowledge gaps. The activity began with a short discussion of tools and techniques already in place for dealing with data. A longer discussion followed of emerging and prospective factors that could have a major influence on data quality in the future. The results of this exercise are summarized in the following section, with some context added for terms that were discussed but not fully elaborated.

In Place

The group identified a number of tools and techniques for dealing with data that are already in widespread use. Many of these were described in the results of Activity 1. The primary tools and techniques discussed included the following:

- Data mining algorithms for clustering, classification, association, and anomaly detection
- Text mining tools that help users get order from unstructured data
- Traffic analysis tools that provide information about the origination, quantity, and movement of data—but not necessarily the content
- Automated tools for cleaning and improving data
- The use of the semantic web and semantic modeling

Emerging

The emerging tools and factors identified by the participants can be grouped in three broad categories: analytic methods, provenance, and data mapping.

- **Growth areas for analytic methods**
  - Data, web, and social analytics – While work is already being done in these areas, most consider these to be very promising fields still in their adolescence.
  - Lexical link analysis (LLA) – LLA combines data mining with network analysis. Expanded use of LLA will help dynamically identify, assess, and predict trends and patterns in data.
  - Artificial Intelligence (AI) agents filtering data – AI agents help with the problem of information overload, and their use is expected to grow as data proliferates. Applications of AI filtering agents include personalized information management, electronic, and management of complex commercial and industrial.[6]
  - Sequence mining – The use of sequence mining, or finding statistically relevant patterns among data examples where the values are delivered in a sequence, are useful for finding temporal sequences in scalable data.
  - Model-based reasoning – Model-based reasoning is being applied in new fields of study and models are becoming more accurate and useful.

[6] http://groups.engin.umd.umich.edu/CIS/course.des/cis479/projects/FISA.html

- **Related to provenance**
  - Models and techniques for recording provenance – A growing number of datasets are available for public use, and users need to be able to apply their own metrics to determine if the data are acceptable. Uniform models and automated techniques for recording provenance will provide a detailed history of the data that users need.
  - Data production maps – Recent research in information quality shows the benefits of managing information as a product. Data production maps have been identified as a critical tool for managing the entire information manufacturing system [Davidson 2004].
- **Related to data mapping**
  - ETL and schema mapping tools – "ETL" refers to the extract-transform-load facilities used for moving bulk data. Manually maintaining the mappings required for this is increasingly impractical as data sets become larger and change more often. Tools are being developed to automatically adapt mappings as schemas evolve.

Prospective

The discussion of tools and factors that might prove useful in the future centered around the need for new or expanded analysis techniques and tools to extend capabilities currently in place. Participants identified new analytic methods, discussed the benefits of quality by design, and expanded their ideas on provenance and data governance.

- **New analytic methods**
  - Markov logic –The use of Markov logic is expected to expand in many fields. Some applications include social network analysis and link prediction, entity resolution, information extraction, link-based clustering, and semantic network extraction from text [Domingos 2009].
  - Formal verification techniques – Formal verification refers to proving or disproving the correctness of intended algorithms underlying a system using formal methods of mathematics. While often used in hardware, the use of formal verification in the software industry is a growth area.
- **Design quality in**

  The concept of "quality by design" has been applied in many fields and is extending to the field of information quality. The quality of data that people use is strongly influenced by the quality of the schema that dictates the structure of the data. Data requirements for quality assurance need to be considered at the time a database schema is designed.
- **Provenance-related**

  The importance of provenance is well recognized. The new focus on the collection of provenance will likely result in a provenance overload. New tools are needed for extracting and harvesting provenance, and also for compression and optimization. The relationship between quality and provenance also needs to be further defined—we know a connection exists, but how do we derive quality from provenance?

- **Data governance-related**

  Many opportunities exist for improving data governance. Automated solutions are in increasing demand, as is the expansion of metadata to include all operational definitions (i.e., record everything that is important to the organization). The integration of data management techniques for structured and unstructured data is also a topic of research.

4 Working Group Proposals

During the late morning of Day 3, workshop participants formed three groups based on their areas of expertise and interest, as demonstrated through their presentations on days 1 and 2. Each of the groups selected topics from the ideas about technology trends identified in workshop activity 2.

Each group worked separately to identify and characterize research proposals by addressing the Heilmeier questions:

1. What are you trying to do? Articulate your objectives using absolutely no jargon. What is the problem? Why is it hard?
2. How is it done today, and what are the limits of current practice?
3. What's new in your approach and why do you think it will be successful?
4. Who cares?
5. If you're successful, what difference will it make? What impact will success have? How will it be measured? [Heilmeier 1991]

The groups worked over the course of approximately 2.5 hours to generate their results and then reconvened for report-out. Although the teams used the Heilmeier questions to guide their work, the primary output of this exercise was a definition of the problem space. Not all questions could be fully answered in the time available. The work was documented on a series of Powerpoint slides that were presented as an outbrief to all workshop participants. The five proposals, as developed by the groups, are listed below.

4.1 Proposal 1

Reduce complexity in order to increase integrity of the information systems to enable effective coordination of multiple data sources and improve agility between humans and systems.

Proposal team

Peter Aiken, David Hay, Douglas MacKinnon, and David Schlesinger

Starting point

The proposal was formulated based on ideas formulated earlier during the workshop:

- Synchronization of data in space and time according to DoD needs (producing systems that allow this)
- Full integration of structured and unstructured data management
  (identify the optimal structure to integrate structured and unstructured data)

Integrating "in place" and "emerging" tools/results/products into "prospective"
(All tool development should take place within an integrated framework or we risk developing stovepiped tools capable of solving problems locally instead of globally.)

What are you trying to do?

- Reduce complexity in order to increase integrity of the information systems to enable effective coordination of multiple data sources.
- Improve agility between humans and systems.

What is the problem?

- Poor interaction between users and subject databases.
- Systems are slow and users can't find the information they need.
- Fragmentation of data across multiple systems causes unacceptable data gaps and errors.

Why is it hard?

- Too many systems developed over time without high-level guiding architecture created disparate/uncoordinated data definitions, quality control rule sets, and subject matter vocabularies.
- The volume of work to correct this individually is overwhelming – an implementable, programmatic solution is required.

How is it done today, and what are the limits of current practice?

- Systems were developed in response to (at the time) individual problems using different technologies. This means that communication is hard and that the resulting brittle system architectures cannot respond to change.
- Individual solutions are developed for individual problems. The resulting myopic solutions are not:
  - scalable
  - transferable
  - able to be applied outside of the solution domain
- Heroic efforts are celebrated as "accomplishments" instead of mourning the lost opportunities that would have contributed to global solutions.

What's new in your approach?

The three-fold approach is:

1. Develop a unifying set of standard patterns to use as the basis for each organization's development of an essential architectural model.
2. Help organizations develop individual models guided by the standard patterns.
3. Capture the various DoD semantic vocabularies (glossary, business rules, interchange formats, ontologies/taxonomies, data/information architecture).

Why do you think it will be successful?

Success is:

- People who use DoD systems will only have a consistent view of the DoD data.
- This will save money.
- This will improve data quality.

- All the information systems in the DoD will be able to accurately, automatically, and instantly exchange information.

Who cares?

This table lists the roles that would care and why.

Role	Why they care
Users and Analysts	• Less time integrating data and more time analyzing it • Better answers
Decision makers	• More trustworthy data will be available • Will receive complete data more rapidly
Project funders	• Transparency improved • More predictable project costs
Commanders	• Better battlefield information more rapidly • More integrated (structured and unstructured information)
System Developers	• Better quality, more coherent specifications

If successful, what difference will it make?

DoD will be able to respond with greater agility to emerging challenges. Impact includes:

- Improved data quality
- Improved decisions
- More trustworthy responses
- More comprehensive information views

How will success be measured?

- Lower costs
- Increased response speed
- More trusted responses
- Faster decision making

4.2 Proposal 2

Create high-value information by automating the integration of several data sources.

Proposal Team

Diana Angelis, Maureen Brown, Tim Menzies, John Talburt, and Richard Wang

What is the problem?

How to guarantee that the quality of the integration has at least as much quality as any one source.

Why is it hard?

The inaccuracy, incompleteness, inconsistency, and other data quality issues in sources; the ontology mismatch between sources; volume of data; and the need to synchronize by time, location, or other parameters.

How is it done today, and what are the limits of current practice?

For some subfields there are good solutions, but in most cases there is not a general solution or an effective tool. This makes each integration a custom, highly-manual process.

What are the limits to current practice?

Scalability, no general solution to ontology mapping.

What's new in your approach?

The automation of the integration.

Why do you think it will be successful?

Because in some small number of limited contexts a highly automated solution has been demonstrated.

Who cares?

Anyone putting together information from multiple systems, those concerned with situational awareness or layered sensors, or anyone trying to get added value from a network.

If successful, what difference will it make?

- Improved defense
- Reduced cost of integrating next generation systems

How will success be measured?

By benchmarking automated integration time, effort, and outcomes against current practice.

4.3 Proposal 3

Develop models, tools, and methods for designing data quality into software over its life cycle, from specification through implementation. A calculus for data quality that captures the uncertainty in the output space would also be developed.

Proposal Team

Diana Angelis, Maureen Brown, Tim Menzies, John Talburt, and Richard Wang

What is the problem?

Current practices in software development and maintenance do not directly address the issue of data quality. They assume that data are correct and there is little ability to signal the occurrence of errors in data or correct them. Current practices focus on the quality of data at the input space. Decision makers need to know the uncertainty of information outputs.

Why is it hard?

Software developers have traditionally assumed that data quality would be addressed by different organizational roles, apart from the development process. Tools and models for evaluating and embedding data quality capability into software have not been developed. Concepts and tools for understanding the uncertainty around the quality of data outputs are not well developed, nor is that information well reported.

How is it done today?

Data quality is addressed entirely apart from software develop process and requires an intensive, time-consuming, and primarily manual assessment and correction process. Current software does not report on the impact of data quality uncertainty on the decision making process.

What are the limits of current practice?

A labor intensive, iterative process performed after the initial specification and development that involves running test data sets, manual auditing and evaluation of output data, determination of root causes of errors, and software change specifications to address error production. Indicators of input and output data quality are not automatically reported.

What's new in your approach?

Data quality will be addressed as part of the software design process and throughout the software life cycle and not as a separate process.

Why do you think it will be successful?

Most data quality issues are already addressed by ad hoc software tools and systems. It should be possible to translate these ad hoc processes into specifications for the original system. There is already a large body of knowledge about designing quality software and an accepted capability and maturity model for software development.

The results of this research can be incorporated into the body of knowledge and practice for software development. Data quality calculus based on a calculation of data quality uncertainty will provide decision makers with the impact of data quality on output information. This new approach will also monitor the quality and uncertainty of data quality during information production.

Who cares?

Everyone concerned with data quality and the usefulness and reliability of information.

If successful, what difference will it make?

Assurance that software systems will produce high-quality information and facilitate understanding of the effect of poor quality on decisions.

What impact will success have?

Decrease the time and effort to assess and improve data quality and significantly reduce the number of data errors produced by software systems.

How will success be measured?

Benchmark against software systems that do not incorporate data quality design.

4.4 Proposal 4

Use provenance information for assessing and improving quality.

Proposal 4 Team

Jon Agre, Susan Davidson, Alan Karr, and Sudha Ram

Background

Provenance is information that describes the life cycle of data and can be used for assessing data quality; attribution; authentication; understanding and replicating process (replication recipe); chain of custody; attribution; digital rights management; audit trails; means of access control; and monitoring.

Quality can be evaluated in many different ways, and provenance is input that can be used for evaluating quality: provenance is an input and quality assessment is the output. So if you have a way of capturing provenance you can use your own technique to browse provenance and give an assessment of quality according to your own beliefs.

Why is it hard?

There are two top-level issues that need to be addressed:

1. How to model, extract, and manage provenance?
2. How to effectively use provenance for assessing provenance?

Detailed questions include:

- How to discover provenance (e.g., from unstructured data)?
  (Provenance can be large—Thousands of times larger than the actual data. Many issues are involved in managing provenance: the level of granularity at which to capture provenance, determining how to aggregate and efficiently store provenance, how to index, and so forth.)
- Who can see what provenance information and under what conditions (interactions with privacy)?
- What is the query language; how to focus attention on important parts of provenance information?
- How can provenance be used to assess and improve data quality (provenance analytics)?

How to reason and reduce volume, yet get enough information from provenance data? Reasoning on provenance (especially real time reasoning) is hard.

How is it done today?

In the DoD:

> Data is being designated as authoritative or not based on the source of the data; when data is entered in a database the source must be recorded (designated as part of the schema).

In the private sector:

> Some provenance models have been proposed and are getting buy-in from the community (e.g. Sudha Ram's W7, Open Provenance Models, ISO8000).

Some workflow systems are being developed that automatically capture workflow provenance (e.g. VisTrails, Kepler, Taverna); prototype database systems are being developed to capture and manage database provenance. Provenance is also being captured at the OS level (e.g. PASS project).

What are the limits of current practice?

Not all provenance information is being tracked; provenance is much broader than just the source of the data.

The authoritative source: how is it being designated?

Methods for stewardship and designating authoritative sources do not scale, and they are time-consuming and person-intensive. Authoritativeness is also a binary decision; there are no intermediate levels of trustworthiness.

We also need feedback to help improve the quality of a source (e.g., checking whether or not the authoritative source was wrong or right after the fact).

What's new in your approach?

This research broadens the definition of provenance beyond the source and proposes to develop new techniques to record, extract, and manage provenance. New techniques to reason about and use provenance will also be developed, along with tools and techniques to use provenance to determine and improve data quality.

Who cares?

Anyone who makes decisions (e.g., in situational awareness, battlefield operations, financial resource allocation, healthcare).

If successful, what difference will it make?

Are people able to make faster, better decisions when provenance is taken into account? (OODA loop: observe, orient, decide, act)

Have we reduced the number of failures based on inaccurate provenance information? (To do this, we need to understand what the characteristics are of failures due to lack of provenance.)

Are we better able to quantify the authoritativeness of data sources using provenance? Are we able to reduce the time/effort to recognize the authoritativeness of a data source?

Can you use feedback to improve the quality of sources felt to be non-authoritative; can you improve decisions by taking more kinds of provenance into account.

Assuming that augmented provenance management is put in, what is the effectiveness of decisions made without provenance versus those that take provenance into account?

How will success be measured?

Other outcomes include detecting "unusable" data sources: we may find that there are data sources that are not useable judging from the provenance of data that it contains.

We need to use a step-wise approach: Since provenance can be huge, we need to stage the level of detail at which it is collected, determine what helps and what doesn't help. What is most important at the next step in provenance?

4.5 Proposal 5

Cost benefit tradeoff (disks may be cheap but people are not cheap).Guarantee privacy while maximizing data quality.[7]

Proposal Team

Jon Agre, Susan Davidson, Alan Karr, and Sudha Ram

What is the problem?

The thesis is that what you do to protect privacy tends to deteriorate data quality. Techniques that work for official statistics data do not necessarily work for other applications. There are different measures of privacy and utility.

Scenario 1: DoD releases data to contractors and must guarantee certain privacy concerns while maximizing the utility of the data for the contractor (i.e. enable the contractor to do as much useful work as possible).

Scenario 2: Provenance captures intermediate data and behavior of processes that may be confidential or proprietary. In a workflow, intermediate data (e.g. medical records, financial data) could be confidential, and modules could be proprietary. Capturing complete provenance information may reveal behavior as well as intermediate data.

Additional questions

Are privacy considerations going to affect the DoD's ability to collect data?

1. The DoD wants to prevent inadvertent release of data that they want to be private. They don't want people snooping external sources (e.g. social networking sites) and figuring things out that are inferences from the data being released.

   For example (covert channels): determining the timing of an invasion in Iraq based on number of late-night pizza orders coming out of the Pentagon.

[7] This proposal was partially addressed due to time limitations.

2. When is it OK to cross-link external and internal data sources and maintain privacy?

   Will putting together heterogeneous data collections enable more information than was possible before?

DoD wants to prevent inadvertent release of data that they want to be private. They don't want people snooping external sources (e.g. social networking sites) and figuring things out that are inferences from the data being released.

What are the limits of current practice?

Externally released data: PII is removed. Understand whether existing techniques developed in other contexts are adequate for the DoD?

5 Research Recommendations Related to Data Quality, Data Analysis, and Data Use

After the workshop, the SEI used the proposals generated by the working groups to craft three recommendations for further research. While these recommendations retain the spirit and intent of the working groups, the following changes were made:

1. Similar proposal elements were combined to eliminate overlap.
2. Further elaboration and context were included when possible.
3. The proposal language was edited for clarity.

5.1 Recommendation 1: Enable the effective integration of data from multiple and disparate data sources.

DoD information systems have been developed to respond to specific problems and use and produce data specific to that system. Because there is no overriding architecture to guide the development of these systems, integrating and coordinating the data from multiple sources is difficult or impossible. Fragmentation of data across multiple systems causes unacceptable data gaps and errors due to uncoordinated data definitions, lack of synchronization, disparate quality control sets, and inconsistent subject matter vocabularies. The volume of work and the amount of time necessary to overcome these limitations is overwhelming.

There is every reason to expect that developing well-integrated DoD information systems would result in important benefits. Systems would be able to automatically and instantly exchange information. Analysts would have a consistent, trustworthy, and comprehensive view of the data, leading to faster responses, lower costs related to finding and gathering information, and improved data for decision making.

Research into the following three areas was recommended as important for achieving unification of data sources:

1. A set of standard patterns should be developed to serve as the basis for the development of each organization's essential architectural model for information. Once a pattern is approved, it can be used by itself or in combination with other patterns in the design of physical databases and implementation of feature-level metadata.
2. Semantic vocabularies (e.g., data architecture, information ontology and taxonomy, data definitions, business rules, interchange formats, glossary, etc.) should be developed for use by organization's sharing data as well as system developers.
3. Tools and methods for utilizing the patterns and vocabularies for system and data specification should be developed. Further, tools should also be developed for the analysis of such specifications and to assure that the integrated data retains the quality level of the constituent sources.

An additional research area related to this recommendation was also noted concerning the current state of data and efforts to migrate it to a standard pattern. For example, two or more records being integrated might contain information on the same real-world entity, but have no unique iden-

tifiers to show that they are related. Or, related records may also contain conflicting information. Entity resolution and other integration difficulties might be partially overcome through the application of new algorithms and techniques.[8] Solutions to entity resolution problems have been proposed that integrate previous approaches based on Markov logic.[9] Markov logic is a relatively new framework for combining first-order logic and probabilistic graphical models.

5.2 Recommendation 2: Employ provenance analytics to ensure data quality to support mission success.

Information provided and used by the DoD is typically assumed to be authoritative even though the origin of the data and its transformations are unknown. Data provenance can be used to ensure that users of data understand important background aspects, including its origin, who or what process created the data, how it was transformed, and any other conditions used to generate the data provided to users. Data provenance has the potential to be a powerful mechanism for understanding data quality, both objectively and subjectively.

Emerging approaches to provenance have focused mainly on metadata about the source of the data. Some tools and techniques address this part of the problem today. A detailed history of the data could be provided to users through the use of uniform models and automated techniques for recording provenance. A new area of provenance analytics could be developed to provide summaries describing "fitness for use" as well as in-depth analysis identifying data quality problems and opportunities for improvement. Further, insight into provenance would enhance a user's ability to quickly judge the authority of data, thereby enabling rapid decision making. While rapid decision making is desirable in any environment (e.g., project planning, financial resource allocation, healthcare), it is particularly crucial for battlefield operations where situational awareness is key to mission success and the safety of our warfighters.

Some of the challenges that will be encountered in this area include the following:

- The volume of information needed to record provenance is potentially large; what is the proper level of granularity to capture and how can provenance information be efficiently stored, aggregated, and indexed?
- How can security/privacy of data be protected in the presence of amalgamated provenance information?

The proposed research in this area recommends focusing on new techniques to model, record, extract, and manage data provenance as data flow within and across systems and enterprises. In particular, techniques to reason about and use provenance should be explored, including the development of tools that use provenance information to determine data quality and how to improve it.

[8] Some open problems and challenges in entity resolution are described in "Entity Resolution: Overview and Challenges" [Garcia-Molina 2004].

[9] "Entity Resolution with Markov Logic" discusses an integrated solution to entity resolution problems based on Markov logic, and describes how previous approaches can be combined effectively [Singha 2006].

5.3 Recommendation 3: Develop models, methods, and tools that support data quality by design during software development.

When data quality is addressed in isolation from the development of the software used by stakeholders to collect, process, analyze, and report it, the value and benefits of the data and the enterprise's resultant information products can suffer. Missed opportunities, bad analysis, and incorrect decisions can all result from a failure to integrate data quality with software requirements. Software developers may assume that data quality considerations have been addressed by other stakeholders such as business analysts or data architects, or they may make wrong assumptions regarding data quality and associated requirements. This can lead to data of unknown quality as well as inefficient and ineffective use of the organization's data. The result is a decision making process fraught with risk due to unknown or poor data quality and missed opportunities to leverage the data. The anticipated benefit to be realized is an increased reliability of information produced by software applications, leading to improved decision making.

The research topic proposed is an exploration of ways to formally address data quality as part of the software development life cycle. Data quality requirements must be considered during the definition of software requirements, specification, and design. Systematic methods must be developed to refine high-level, abstract data quality goals into concrete data quality requirements. This includes methods to analyze and preserve the integrity of data quality requirements throughout the software development process. Specifically noted was the idea for a data quality calculus that would be used to characterize data quality uncertainty and provide confidence information for decision makers utilizing data from the system as well as to analyze the impact and trade-offs among system and data architecture and design decisions.

Appendix: Workshop Agenda

AGENDA

Issues and Opportunities for Improving the Quality and Use of Data in the DoD

Workshop Location	
Software Engineering Institute NRECA Building, Suite 200 4301 Wilson Boulevard Arlington, VA 22203	Training Room B Second Floor Tel: +1 703.908.8200

Dates of Workshop	Start Time	End Time
October 26-28, 2010	8:00 a.m. October 26	5:00 p.m. October 28

Facilitator	Documentarian
Mark Kasunic	Erin Harper

Presenters	
Nabil R. Adam	U.S. Department of Homeland Security
Peter Aiken	Data Blueprint, Inc.
Diana I. Angelis	Naval Postgraduate School
Mary Maureen Brown	University of North Carolina at Charlotte
Susan B. Davidson	University of Pennsylvania
Brett Dorr	DataFlux
Robert Flowe	OUSD (AT&L)[†]
David C. Hay	Essential Strategies, Inc.
John Horst	National Institute of Standards and Technology (NIST)
Alan Karr	National Institute of Statistical Sciences (NISS)
David Loshin	Knowledge Integrity, Inc.
Douglas J. MacKinnon	Naval Postgraduate School
Tim Menzies	West Virginia University
Sudha Ram	University of Arizona
Thomas C. Redman	Navesink Consulting Group, LLC
David Schlesinger	Metadata Security, LLC
John R. Talburt	University of Arkansas at Little Rock

[*] *OUSD (AT&L) is Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics.*

AGENDA

Attendees

Jon Agre	Institute for Defense Analyses
Anita Carleton	Software Engineering Institute
Erin Harper	Software Engineering Institute
David Jakubek	OUSD (AT&L), DDR&E[†]
Mark Kasunic	Software Engineering Institute
Patricia Lothrop	OUSD(AT&L)
Michael May	OUSD (AT&L), DDR&E
Ryan Mckenzie	Amentra, Contract Support to OUSD(AT&L)
Richard Wang	Headquarters, Dept. of the Army
David Zubrow	Software Engineering Institute

[†] *DDR&E is Director, Defense Research & Engineering.*

Workshop Ground Rules

1. The facilitator provides process. Feel free to offer input and process checks.
2. One speaker at a time. It helps us all to hear and understand (please keep offline or cross-talk to a minimum).
3. The facilitator will strive for an inclusive, participative process—for extroverts and introverts alike.
4. Stay on topic and be concise. It helps us achieve our goals.
5. The facilitator will manage our time for efficiency and productivity. Please understand the need to keep us moving ahead.
6. This workshop is interactive and the objective is to obtain ideas and feedback from the invited presenters. For this reason, a condition of the daily honorarium is full attendance on all days of the workshop.
7. Let's keep focused—no cell phones, email or surfing during the sessions.

AGENDA - DAY 1

Tuesday, October 26

Start	End	Dura-tion	Who?	Topic
8:00 a.m.	8:15 a.m.	0:15	Mark Kasunic Michael May David Zubrow	Welcome & Opening Remarks
8:15 a.m.	8:45 a.m.	0:30	All	Introductions
8:45 a.m.	9:30 a.m.	0:45	Mark	Agenda Review, Ground Rules & Logistics
9:30 a.m.	10:05 a.m.	0:35	Robert Flowe	OSD AT&L Perspectives & Initiatives: Acquisition Visibility and Data Quality
10:05 a.m.	10:20 a.m.	0:15	All	Break
Data Quality Issues in DoD Acquisition				
10:20 a.m.	10:55 a.m.	0:35	Mary Maureen Brown	Resource Management Decision 700 & Programmatic Interdependencies
10:55 a.m.	11:30 a.m.	0:35	Diana Angelis	Measuring Transaction Costs in DoD Acquisition Programs
11:30 a.m.	12:00 p.m.	0:30	All	Key Take-Aways
12:00 p.m.	1:00 p.m.	1:00	All	Lunch
Managing Data Quality				
1:00 p.m.	1:35 p.m.	0:35	Thomas Redman	Toward a Management System for Data
1:35 p.m.	2:10 p.m.	0:35	Peter Aiken	Practical Considerations for Rapidly Improving Quality in Large Data Collections
2:10 p.m.	2:45 p.m.	0:35	David Hay	How to Address Data Quality in the Department of Defense
2:45 p.m.	3:00 p.m.	0:15	All	Break
3:00 p.m.	3:30 p.m.	0:30	All	Key Take-Aways
Relationship of Data Quality & Data Security				
3:30 p.m.	4:05 p.m.	0:35	David Schlesinger	Methods to Engineer Information Quality and Protection into Opera-

				tional Processes
4:05 p.m.	4:40 p.m.	0:35	Nabil Adam	Security and Confidentiality in Out-sourced Databases
4:40 p.m.	5:00 p.m.	0:30	All	Key Take-Aways
				Adjourn for the Day

AGENDA - DAY 2

Wednesday, October 27

Start	End	Duration	Who?	Topic
8:00 a.m.	8:15 a.m.	0:15	Mark	Recap, Check In, Adjust
Understanding the Quality of Real-Time Data				
8:15 a.m.	8:50 a.m.	0:35	Douglas MacKinnon	Issues and Opportunities for Improving the Quality and Use of Data Within DoD
8:50 a.m.	9:25 a.m.	0:35	John Horst	Achieving Information Quality
9:25 a.m.	10:00 a.m.	0:35	John Talburt	Information Quality Education and Research at UALR
10:00 a.m.	10:15 a.m.	0:15	All	Break
10:15 a.m.	10:45 a.m.	0:30	All	Key Take-Aways
Data Quality Monitoring and Impact Analysis				
10:45 a.m.	11:20 a.m.	0:35	David Loshin	Evaluating the Business Impacts of Poor Data Quality
11:20 a.m.	11:55 a.m.	0:35	Brett Dorr	Data Quality, Logistics, and USTRANSCOM
11:55 a.m.	12:15 p.m.	0:20	All	Key Take-Aways
12:15 p.m.	1:15 p.m.	1:00	All	Lunch
1:15 p.m.	1:50 p.m.	0:35	Susan Davidson	Data Provenance: The Foundation of Data Quality
1:50 p.m.	2:25 p.m.	0:35	Sudha Ram	Determining Data Quality Based on Provenance
2:25 p.m.	2:45 p.m.	0:20	All	Key Take-Aways
2:45 p.m.	3:00 p.m.	0:15	All	Break
Modeling Data Quality				
3:00 p.m.	3:35 p.m.	0:35	Tim Menzies	Selecting Quality Data
3:35 p.m.	4:10 p.m.	0:35	Alan Karr	Data Quality Research that Builds on Data Confidentiality
4:10 p.m.	4:45 p.m.	0:35	All	Key Take-Aways
4:45 p.m.	4:30 p.m.	0:15	Mark	Preview of Day 3 Activities
				Adjourn for the Day

AGENDA - DAY 3

Thursday, October 28

Start	End	Duration	Who?	Topic
8:00 a.m.	8:15 a.m.	0:15	Mark	Recap, Check in, and Set-up
8:15 a.m.	8:30 a.m.	0:15	Mark	Context Map - Introduction
8:30 a.m.	10:00 a.m.	1:30	All	Develop the Context Map
10:00 a.m.	10:15 a.m.	0:15	All	Break
10:15 a.m.	10:30 a.m.	0:15	Mark	Waves of Innovation - Introduction
10:30 a.m.	11:15 a.m.	0:45	All	Develop the Waves of Innovation Chart
11:15 a.m.	11:30 a.m.	0:15	Mark	Break-Out Groups: Set-up
11:30 a.m.	12:30 p.m.	1:00	All	Lunch
12:30 p.m.	1:30 p.m.	1:00	All*	Develop Research Agenda – Five-Year Horizon
1:30 p.m.	2:15 p.m.	0:45	All	Present Results: "Five Year Horizon" Research Agenda
2:15 p.m.	2:45 p.m.	0:30	All	Recommendation roll-up: Multivote AND break
2:45 p.m.	3:30 p.m.	0:45	All*	Develop Research Agenda – Ten-Year (+) Horizon
3:30 p.m.	4:15 p.m.	0:45	All	Present Results: " Ten-Year (+) Horizon" Research Agenda
4:15 p.m.	4:25 p.m.	0:10	All	Recommendation Roll-Up: Multivote
4:25 p.m.	4:40 p.m.	0:15	Mark	Summarize Results
4:40 p.m.	5:00 p.m.	0:20	All	Next Steps
				Adjourn

*** Conducted in subgroups.**

References/Bibliography

URLs are valid as of the publication date of this document.

[Davidson 2004]
Davidson, Bruce; Lee, Yang; and Wang, Richard. "Developing Data Production Maps." *International Journal on Healtcare Technology and Management* 6, 2 (2004).

[Domingos 2009]
Domingos, Pedro and Lowd, Daniel. *Markov Logic: An Interface Layer for Artificial Intelligence.* Morgan and Claypool Publishers, 2009.

[English 2009]
English, Larry P. *Information Quality Applied: Best Practices for Improving Business Information, Processes and Systems.* Indianapolis, IN.: Wiley, 2009.

[GAO 2009]
United States Government Accountability Office (GAO). “DoD Business Systems Modernization: Recent Slowdown in Institutionalizing Key Management Controls Needs to Be Addressed,” (GAO-09-586), 2009.

[Garcia-Molina 2004]
Garcia-Molina, Hector. "Entity Resolution: Overview and Challenges," *Lecture Notes in Computer Science 3288* (2004).

[Heilmeier 1991]
Heilmeier, George H. "An Interview with George H. Heilmeier," Arthur Norberg, interviewer. Charles Babbage Institute, University of Minnesota, Minneapolis, March 27, 1991.

[Power 2010]
Power, Dan. "Taxonomies and the Semantic Web in Product Master Data." *Information Management*, Jan/Feb 2010.

[Singha 2006]
Singla, Parag and Domingos, Pedro. "Entity Resolution with Markov Logic." *Proceeding ICDM '06 Proceedings of the Sixth International Conference on Data Mining*, IEEE Computer Society, Washington, DC, 2006.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
(Leave Blank)	March 2011	Final

4. TITLE AND SUBTITLE	5. FUNDING NUMBERS
Issues and Opportunities for Improving the Quality and Use of Data in the Department of Defense	FA8721-05-C-0003

6. AUTHOR(S)

Mark Kasunic, David Zubrow, Erin Harper

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Software Engineering Institute Carnegie Mellon University Pittsburgh, PA 15213	CMU/SEI-2011-SR-004

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSORING/MONITORING AGENCY REPORT NUMBER
HQ ESC/XPK 5 Eglin Street Hanscom AFB, MA 01731-2116	

11. SUPPLEMENTARY NOTES

12A DISTRIBUTION/AVAILABILITY STATEMENT	12B DISTRIBUTION CODE
Unclassified/Unlimited, DTIC, NTIS	

13. ABSTRACT (MAXIMUM 200 WORDS)

The Department of Defense (DoD) is becoming increasingly aware of the importance of data quality to its operations, leading to an interest in methods and techniques that can be used to determine and improve the quality of its data. The Office of the Secretary of Defense for Acquisition, Technology, and Logistics (OSD [AT&L]), Director, Defense Research & Engineering (DDR&E) sponsored a workshop to bring together leading researchers and practitioners to identify opportunities for research focused on data quality, data analysis, and data use. Seventeen papers were accepted for presentation during the workshop. During workshop discussion participants were asked to identify challenging areas that would address technology gaps and to discuss research ideas that would support future DoD policies and practices. The Software Engineering Institute formed three primary recommendations for areas of further research from the information produced at the workshop. These areas were integrating data from disparate sources, employing provenance analytics, and developing models, methods, and tools that support data quality by design.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Data quality, workshop report, information quality, provenance	56

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL